applied sciences

MDPI

Article

Breast Cancer Prediction Based on Differential Privacy and Logistic Regression Optimization Model

Hua Chen [1,*], Nan Wang [1], Yuan Zhou [1,2], Kehui Mei [1], Mengdi Tang [1] and Guangxing Cai [1]

[1] School of Science, Hubei University of Technology, Wuhan 430068, China; m15271885698@163.com (N.W.); 13207169394@163.com (Y.Z.); 102112265@hbut.edu.cn (K.M.); m15527920661_1@163.com (M.T.); slxcai1964@163.com (G.C.)
[2] School of Computer Science and Technology, Wuhan University of Bioengineering, Wuhan 430415, China
* Correspondence: 20070002@hbut.edu.cn

Abstract: In order to improve the classification effect of the logistic regression (LR) model for breast cancer prediction, a new hybrid feature selection method is proposed to process the data, using the Pearson correlation test and the iterative random forest algorithm based on out-of-bag estimation (RF-OOB) to screen the optimal 17 features as inputs to the model. Secondly, the LR is optimized using the batch gradient descent (BGD-LR) algorithm to train the loss function of the model to minimize the loss. In order to protect the privacy of breast cancer patients, a differential privacy protection technology is added to the BGD-LR model, and an LR optimization model based on differential privacy with batch gradient descent (BDP-LR) is constructed. Finally, experiments are carried out on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Meanwhile, accuracy, precision, recall, and F1-score are selected as the four main evaluation indicators. Moreover, the hyperparameters of each model are determined by the grid search method and the cross-validation method. The experimental results show that after hybrid feature selection, the optimal results of the four main evaluation indicators of the BGD-LR model are 0.9912, 1, 0.9886, and 0.9943, in which the accuracy, recall, and F1-scores are increased by 2.63%, 3.41%, and 1.76%, respectively. For the BDP-LR model, when the privacy budget ε is taken as 0.8, the classification performance and privacy protection effect of the model reach an effective balance. At the same time, the four main evaluation indicators of the model are 0.9721, 0.9975, 0.9664, and 0.9816, which are improved by 1.58%, 0.26%, 1.81%, and 1.07%, respectively. Comparative analysis shows that the models of BGD-LR and BDP-LR constructed in this paper perform better than other classification models.

Keywords: breast cancer; feature selection; batch gradient descent; differential privacy; logistic regression

Citation: Chen, H.; Wang, N.; Zhou, Y.; Mei, K.; Tang, M.; Cai, G. Breast Cancer Prediction Based on Differential Privacy and Logistic Regression Optimization Model. *Appl. Sci.* **2023**, *13*, 10755. https://doi.org/10.3390/app131910755

Academic Editor: Luis Javier García Villalba

Received: 9 September 2023
Revised: 24 September 2023
Accepted: 26 September 2023
Published: 27 September 2023

1. Introduction

Cancer is the leading cause of human mortality worldwide, and the treatment of cancer consumes a lot of medical resources and increases the burden on society, so that cancer has become a common social issue all over the world [1]. Breast cancer is one of the malignancies with the highest morbidity and mortality in women [2]. The symptoms of early-stage breast cancer are not obvious, and advanced cancer cells will rapidly metastasize, leading to systemic multiorgan lesions that will directly threaten the lives of patients, so early diagnosis is the key to improving the survival rate of breast cancer patients [3].

At present, there are three common methods for early diagnosis of breast cancer: clinical evaluation [4], imaging evaluation [5], and tissue biopsy. Using machine learning methods on these detection data for data analysis and data mining can assist doctors in reducing misdiagnosis and missed diagnosis caused by subjective factors and improve the detection rate of breast cancer [6]. Machine learning is at the core of artificial intelligence and data science. As machine learning continues to be optimized, cancer prediction accuracy continues to improve [7].

Machine learning methods have worked well in the diagnosis of breast cancer. Arpit, B. et al. [8] used the multilayer perceptron (MLP), K-nearest neighbor (KNN), genetic algorithm (GP), and random forest (RF) to classify breast cancer cells, and the experimental results showed that the RF model had the highest classification accuracy, which could reach 0.9624. Fatih Ak, M. [9] used data visualization and multiple machine learning models to analyze breast cancer diagnosis, and experiments showed that the LR model had the highest classification accuracy of 0.981. However, the algorithm is data demanding and does not have the performance to handle missing values. Mahesh, T.R. et al. [10] used six methods including an ensemble learning technique, support vector machine (SVM), KNN, decision tree (DT), RF, and LR to classify and predict breast cancer data, of which the accuracy of the ensemble learning technique was the highest at 0.9814, followed by the LR model at 0.9632. Naseem, U. et al. [11] proposed a breast cancer diagnosis system and prognosis automatic detection system based on an ensemble of classifiers. Experimental results on the WDBC dataset showed that the ensemble method is superior to other single methods, with an accuracy of 0.9883.

Recently, scholars have made a lot of contributions to improving the accuracy of breast cancer prediction, and some of them have improved the machine learning model [12,13]. Wang et al. [14] proposed an SVM-based weighted area under the curve (AUC) ensemble learning model for breast cancer diagnosis, and the results showed that the proposed WAUCE structure could improve the diagnostic accuracy by 0.94% on small datasets, but the WAUCE model has a long computational time. Zheng et al. [15] extracted breast cancer tumor features and diagnosed them according to a K-means and SVM hybrid algorithm, and the results showed that the hybrid algorithm improved the accuracy to 0.9738. Ajay Kumar et al. [16] used a user-defined weighted set voting scheme for breast cancer classification, assigned custom-based weights, and used an ensemble classifier to outperform each estimator for the final classification of cancer. The highest accuracy of the proposed ensemble classifier reached 0.9647. X. Jia et al. [17] proposed a whale optimization-based algorithm to improve the accuracy of breast cancer classification by iteratively adjusting the parameters of SVM. Experiments were carried out on the WDBC dataset, and the results showed that the WOA-SVM model had higher classification accuracy than the traditional breast cancer classification models, with an accuracy of 0.975.

Other scholars focus on the feature selection of data [18–24]. In the process of prediction, feature selection can eliminate irrelevant variables and redundant features to achieve effective dimensionality reduction and improve the accuracy of the algorithm [25]. Rao, H. et al. [26] proposed a new feature selection method based on bee colony (ABC) and extreme gradient boosting (XGBoost), which effectively reduced the dimensionality of the dataset, and the accuracy of the XGBoost was 0.928 on the WDBC dataset. Nevertheless, the study is limited to theories related to decision trees. Alghereiry et al.'s [27] study on the WDBC dataset showed that the LR model was the best classifier, and the accuracy of the LR model could be improved from 0.972 to 0.982 by using the forward feature selection method. Abdel-Basset, M. et al. [28] proposed a new grey wolf optimizer (GWO) algorithm combining two-stage variation to solve the feature selection problem and selected the KNN classifier to classify breast cancer data, of which the accuracy of the final model could reach 0.9482. Mahesh, T.R. et al. [29] proposed a breast cancer prediction XGBoost ensemble model based on known feature patterns. In order to deal with the impact of data balance on classification results, SMOTE was used to process the data, and then the naïve Bayes classifier (NB), DT, and RF were combined with XGBoost to classify the data. According to experimental analysis, the classification effect of the XGBoost–random forest ensemble classifier was the best. The classification accuracy of the model was 0.982. Singh, L.K. et al. [30] proposed a unique feature selection method based on eagle strategy optimization (ESO), the gravitational search optimization (GSO) algorithm, and their hybrid algorithm, which could select the fewest features to achieve the highest accuracy. Experimental results showed that the proposed method achieved great results on the WDBC dataset with an accuracy of 0.9896.

Machine learning methods have contributed to the early diagnosis of breast cancer. Making improvements to the classification model and processing the data using feature selection methods can improve the classification of breast cancer, but the classification of breast cancer is still not optimal for the existing research. In addition, there is a risk of leaking specific private information in the training data to attackers through the structure of the model [31]. In recent years, the leakage of private information of patients has occurred frequently. With the advancement of cloud technology and big data, it is easier for attackers to collect patients' private information and speculate about patients' sensitive information through correlation and other means [32]. Therefore, while combining machine learning models with cancer diagnosis, it is also necessary to pay attention to the privacy protection of data.

Commonly used privacy protection techniques are anonymity-based privacy protection, encryption-based privacy protection, and noise-based privacy protection [33]. Differential privacy technology based on noise was proposed by Dwork [34] of Harvard University in 2006 by adding a series of "noise" to the original data, making it difficult for attackers to achieve accurate calculation of an individual user's privacy data, so as to improve the efficiency of data sharing and use under the premise of protecting data security. The application of privacy protection technology in data mining has become a research hotspot in the field of artificial intelligence [35–37], and how to achieve the combination with machine learning with a smaller accuracy loss cost is still an urgent problem to be solved [38].

In summary, there are two main problems in the study of breast cancer prediction: (1) how to improve the prediction effect through feature selection and model improvement; (2) how to improve the classification effect and at the same time make the model have privacy protection function. To solve these two problems, this paper proposes a new hybrid feature selection method to process data. At the same time, it combines differential privacy technology and a logistic regression algorithm to construct a breast cancer classification model with higher classification performance and data privacy protection. The main process is shown in Figure 1, and the main contributions are as follows:

(1) Improve the effect of breast cancer prediction. Firstly, a new hybrid feature selection method is proposed to eliminate weakly correlated features and redundant features, which is divided into two steps: in the first step, the features with an absolute value of the Pearson correlation coefficient greater than or equal to 0.3 are screened out; in the second step, the optimal combination of features is screened to find the final features by the iterative RF-OOB algorithm. Then, the BGD algorithm is used to optimize the LR, and the loss function of the model is trained to minimize the loss to improve the classification effect of the model. In order to verify the effectiveness of hybrid feature selection, a control group experiment is set up to compare the results.
(2) Differential privacy protection technology is added to the process of breast cancer prediction. In the BGD algorithm, Gaussian noise is added to each layer of gradient descent, which makes the model have accurate classification performance while protecting data privacy. Finally, the optimal results of the model in this paper are compared with the results in other papers.

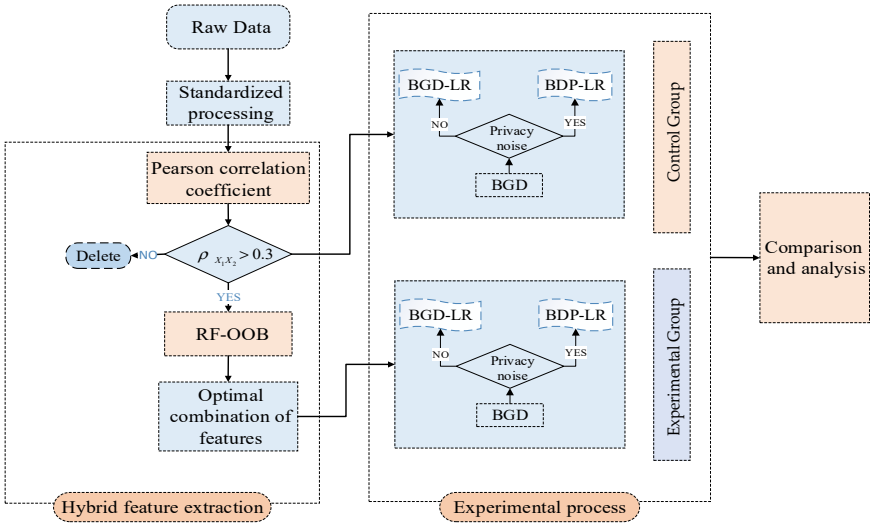

Figure 1. Flowchart of breast cancer classification model with data privacy protection.

2. Methods and Materials

This section introduces the basic theoretical concepts of differential privacy preserving techniques, feature selection methods, LR algorithms, and batch gradient descent algorithms. In particular, the feature selection methods contain the Pearson correlation test and random forest algorithm based on out-of-bag estimation.

2.1. Differential Privacy

Definition 1 (differential privacy). *If there is a mechanism F satisfying differential privacy protection, the sum of its outputs is S and* Pr[] *is the probability of the output results. For all adjacent data sets A and A′, there is*

$$\frac{\Pr[F(A) = S]}{\Pr[F(A') = S]} \leq e^{\varepsilon} \tag{1}$$

Mechanism F is said to satisfy differential privacy [39], where ε is the privacy budget. When ε is smaller, F needs to give a very similar output and therefore provide higher privacy. Conversely, a larger ε allows F to give less similar outputs, providing less privacy. Satisfying Equation (1) is called the strictly satisfying ε -differential privacy definition.

Definition 2 (approximate differential privacy). *In the process of experimentation, because of too-strict protection, the availability of data will be seriously affected. In order to solve this problem, Dwork et al.* [40] *gave the concept of approximate differential privacy, when the*

$$\Pr[F(A) = S] \leq e^{\varepsilon}\Pr[F(A') = S] + \delta \tag{2}$$

where the privacy parameter δ is a small constant indicating the "probability of failure" that does not meet this approximate differential privacy definition, and we set the δ to 0.00001. Like strict differential privacy, approximate differential privacy, which is known as (ε,δ)-differential privacy, also satisfies sequence combinatory and parallel combinatory.

Satisfying approximate difference privacy means that if we change an element in the database, the probability of the output should be close to the probability of the original data, thus protecting the original data from leakage [41].

Definition 3 (global sensitivity). *The sensitivity of the function reflects the degree to which the output changes when the input of the function changes. For a query function* $f : D \to \mathbb{R}^k$ *and a norm function* $\|\cdot\|$*, the sensitivity is*

$$s(f, \|\cdot\|) = \max_{d(A,A')\leq 1} \left\| f(A) - f(A') \right\| \tag{3}$$

The norm function is usually L_1 or L_2 [42], the length of the vector V is k, the L_1 norm is defined as $\|v\|_1 = \sum_{i=1}^{k} |v_i|$, thus the sum of the elements of the vector and the L_2 norm is defined as $\|v\|_2 = \sqrt{\sum_{i=1}^{k} v_i{}^2}$. In two-dimensional space, the L_2 norm is always less than or equal to the L_1 norm.

Definition 4 (Gaussian noise mechanism). *The Gaussian mechanism cannot satisfy strict ε-differential privacy, but it can satisfy (ε,δ)-differential privacy, so for the function* $f : D \to \mathbb{R}^k$ *, the Gaussian mechanism defined below is applied to obtain F(A) satisfying approximate differential privacy:*

$$F(A) = f(A) + N(\sigma^2) \tag{4}$$

where $N(\sigma^2)$ *represents the Gaussian (normal) distribution sampling result with a mean of 0 and a variance of* σ^2*, where* $\sigma^2 \geq c\Delta_2 f^2/\varepsilon^2$*,* $c^2 \geq 2\ln(1.25/\delta)$ *[39],* $\Delta_2 f$ *is the L global sensitivity.*

2.2. Pearson Correlation Coefficient Test

The Pearson correlation coefficient [43] is used to test the correlation of each feature with the target variable, and the Pearson correlation coefficient formula is as follows:

$$\rho_{X_m X_n} = \frac{Cov(X_m, X_n)}{\sqrt{DX_m * DX_n}} = \frac{E(X_m X_n) - EX_m * EX_n}{\sqrt{DX_m * DX_n}} \tag{5}$$

where $\rho_{X_m X_n}$ indicates the correlation coefficient between two variables, $Cov(X_m, X_n)$ indicates the covariance between two variables, EX_m indicates the expectation of the variable, and DX_m represents the variance of the variable.

According to Equation (5), the correlation coefficient between each feature and the target variable is calculated. Based on the thresholds set in this paper, values with absolute values of correlation coefficients greater than the thresholds are filtered to weed out variables other than those with weak correlations. The filtered feature variables are used as candidate features for secondary feature screening.

2.3. Random Forest Algorithm Based on Out-of-Bag Estimation (RF-OOB)

Because of the subset of candidates obtained by the correlation coefficient screening method, features with high correlation will occur, and redundancy between such features will affect the classification results of the model. Therefore, secondary feature screening of random forests estimated outside the bag is also required for candidate subsets.

About 36.8% of the sample data of the RF model was not extracted by the bootstrap sampling [44], which is out-of-bag data of the decision tree. Out-of-bag estimation is to use these data to test the model, and the ratio of misclassified data to the total number of out-of-bag data is out-of-bag estimation, which is also an unbiased estimation of the generalization error of the ensemble classifier. Due to the presence of out-of-bag samples, cross-validation testing is not required for random forest out-of-bag estimation.

As shown in Equation (6), the sum of the out-of-bag score (oob-score) and the out-of-bag error is 1. For a single decision tree T_i trained by the sampling method, operating with

out-of-bag data produces an oob-score. So, for T decision trees, there will be T oob-scores. Finally, the mean is derived to obtain the oob-score for the whole random forest.

$$oob\_score = 1 - \frac{\sum_{i=1}^{n} (f_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y})^2} \tag{6}$$

2.4. Logistic Regression (LR)

LR is a classification algorithm based on logarithmic probability functions. Its core idea is to nest an S-shaped sigmoid function on the basis of linear regression, so as to convert the output result of linear regression into a value close to 0 or 1, and the sigmoid function formula is:

$$g(z) = \frac{1}{1 + e^z} \tag{7}$$

where $z = w^T \cdot x$, w is the weight that needs to be learned, and x is the sample feature vector. $g(z)$ represents the predicted probability value corresponding to the event when the event is inferred from the sample.

The fitting function $H_\theta(x)$ for LR is:

$$H_\theta(x) = g\left(\theta^T x\right) = \frac{1}{1 + e^{-\theta^T x}} \tag{8}$$

where $P(y = 1|x;\theta) = H_\theta(x)$, $P(y = 0|x;\theta) = 1 - H_\theta(x)$.

The loss function for LR is:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log\left(H_\theta(x^{(i)}) + (1 - y^{(i)}) \log\left(1 - H_\theta(x^{(i)})\right)\right)\right] \tag{9}$$

In LR model, the parameters are generally estimated by the maximum likelihood method [45]. The loss function can measure the gap between the actual variable values and the predicted values. The smaller the loss function, the more accurate the predicted values are. In general, if the difference between the loss function values of the training set and the test set is very small, which both achieve low loss values, then the model can be considered to perform well on both the training set and the test set with a good fit [46].

2.5. Batch Gradient Descent (BGD)

Gradient descent is a commonly used optimization algorithm. Its core idea is to gradually adjust the parameters through iteration, so that the loss function of the model reaches the minimum value. The BGD algorithm is a variant form of the gradient descent algorithm.

$$\theta_{n+1} = \theta_n - \nabla \ell(\theta; x^{(i)}, y^{(i)}) \tag{10}$$

where $\nabla \ell(\theta; x^{(i)}, y^{(i)})$ denotes the gradient of the function $\ell(\theta)$ with respect to the parameter θ. The BGD algorithm uses the entire training set for each iteration and computes the local gradient of the error function with respect to the parameter vector θ while proceeding to the next iteration in the direction of the gradient descent until the algorithm converges to a minimum value.

3. Selection of Indicators for the Evaluation

Evaluation indicators are quantitative indicators for evaluating the performance of the model, and if the selected evaluation indicators are not reasonable, it will affect the orientation of the result analysis. Therefore, different evaluation indicators should be selected for specific data and models.

In this paper, breast cancer prediction is a binary classification problem, then the classification results can generate a confusion matrix, as shown in Table 1, where TP indicates that the positive class is predicted as the number of positive classes; TN indicates that the negative class is predicted as the number of negative classes, which can be referred to as the true counterexample; FP indicates that the negative class is predicted as the number of positive classes, which is referred to as the first type of error; FN indicates that the positive class is predicted as the number of negative classes, which is referred to as the second type of error.

Table 1. Confusion matrix.

Label		Predicted Results	
		Positive	**Negative**
Real situation	Positive	TP	FN
	Negative	FP	TN

In order to measure the classification effect of the model in this paper, the four perspectives of the model—the overall accuracy of the model, the accuracy of the positive class prediction, the coverage ability of the positive class, and the comprehensive performance—are taken into account. The accuracy, precision, recall, and F1-score are selected as the four main evaluation indicators to evaluate the prediction effect of the model. And these evaluation indicators have values between 0 and 1. The closer the value is to 1, the better the classification effect of the model is. The receiver operating characteristic (ROC) curve is also selected as an assistant indicator to compare the classification effect among the models. The specific meanings are as follows:

(1) Accuracy: represents the total proportion of all predictions that are correct (positive and negative categories) and can be expressed as $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$. For breast cancer prediction, a high accuracy rate indicates that the model is better at correctly classifying both malignant and benign tumors. Accuracy is justified because it provides an assessment of overall classification accuracy and can help determine the model's ability to discriminate between the two types of tumors.

(2) Precision: indicates how many of the samples predicted to be positive are truly positive, which can be expressed as $precision = \frac{TP}{TP+FP}$, also known as PPV.

(3) Recall: indicates how many positive cases in the sample were predicted correctly, which can be expressed as $precision = \frac{TP}{TP+FN}$, also known as TPR.

(4) $F1$-score: it can be expressed as $F_1 = 2\frac{PPV \times TPR}{PPV+TPR}$. $F1$-score is a comprehensive evaluation index of extrinsic methods. For breast cancer prediction, $F1$-score is reasonable because it balances the model's ability to correctly classify malignant and benign tumors. It also comprehensively evaluates the precision and recall of the model, which is one of the very important evaluation indicators.

(5) The receiver operating characteristic (ROC) curve: in the ROC curve, the horizontal axis is the false positive rate (FPR) and the vertical axis is the true positive rate (TPR). The points closer to (0, 1) correspond to the better classification performance of the model. AUC is the area under the ROC curve, between 0 and 1. As a numerical value it can be visualized to evaluate the classifier, the larger the value the better. When $AUC = 1$, it is a perfect classifier. When $0.5 < AUC < 1$, it is better than random guessing. When $AUC = 0.5$, like random guessing, the model has no predictive value. When $AUC < 0.5$, the model is less predictive than random guessing.

4. Data Preprocessing

Data preprocessing helps to improve the accuracy of analysis results. For different datasets and different tasks, there will be different data preprocessing methods. In this

paper, the WDBC dataset is first introduced in detail, followed by Z-score standardization of the data according to the characteristics of the dataset.

4.1. Introduction to Data

The WDBC dataset used in this paper was provided by the renowned Dr. Williams of the University of Wisconsin Institute for Clinical Medicine [47] and the eigenvalues were computed from digitized images of fine needle aspiration (FNA) of breast masses. The dataset contains 569 sets of experimental samples. The following ten characteristics of the nucleus of the cells taken from each subject are mainly collected: radius, perimeter, smoothness, area, compactness, concavity, symmetry, texture, concave points, and fractal dimension. Of the experimental samples, 357 sets of data are for benign samples of breast cancer and 212 sets of data are for malignant samples of breast cancer. The breast cancer dataset has one sample label (benign and malignant) and 30 features. The first 10 features are the mean values of the nuclei feature values in the sample images, the 11th to 20th features are the standard deviations of the nuclei feature values, and the 21st to 30th features are the maximum values of the nuclei feature values. The classification label represents the type of breast cancer.

4.2. Data Standardization

Some of the feature data of the WDBC dataset are shown in Table 2, from which it can be seen that there are differences in the magnitude of each feature, and if not standardized, direct experiments will lead to the inability to obtain the real results of the research object. In order to reduce the impact of the data dimension on the model, the data need to be processed dimensionlessly. Commonly used dimensionless processing methods are min–max (normalization) and Z-score standardization [48]. In this paper, based on the characteristics of the WDBC dataset, Z-score standardization is applied to the data.

Table 2. Partial sample characteristic data.

Texture_Mean	Perimeter_Mean	Area_Mean	Smoothness_Mean	Symmetry_Mean
10.38	122.8	1001	0.1184	0.2419
17.77	132.9	1326	0.08474	0.1812
21.25	130	1203	0.1096	0.2069
20.38	77.58	386.1	0.1425	0.2597
14.34	135.1	1297	0.1003	0.1809
15.7	82.57	477.1	0.1278	0.2087
19.98	119.6	1040	0.09463	0.1794

5. A Logistic Regression Optimization Model Based on Hybrid Feature Selection and Differential Privacy

In order to improve the effectiveness of the breast cancer classification model and protect the patient's privacy, first, a hybrid feature selection method that can effectively eliminate redundant variables and select the optimal features is proposed. Second, the LR model is optimized using the BGD algorithm to minimize the loss function of the model. Finally, a logistic regression optimization model based on the hybrid feature selection and differential privacy is proposed by adding the Gaussian noise mechanism on this basis.

5.1. Hybrid Feature Selection

In order to improve the model's accuracy, this paper proposes a new hybrid feature selection method, which combines the Pearson correlation test and the RF-OOB algorithm to effectively eliminate irrelevant and redundant features. The method is divided into two parts, as shown in Figure 2: in the first part, the Pearson correlation coefficient is first utilized to measure the correlation between each feature and the target variable, and the k features whose absolute value of the correlation with the target variable is greater than or equal to 0.3 are screened out from the sample training set $D(X_m^n, Y_m)$. In the second part, the

out-of-bag estimation random forest algorithm is used to calculate the feature importance of the remaining k features. Simultaneously, feature combinations are performed according to the feature scores from high to low. Lastly, the feature combinations with the highest score are iteratively filtered to obtain k' features to realize redundant feature removal.

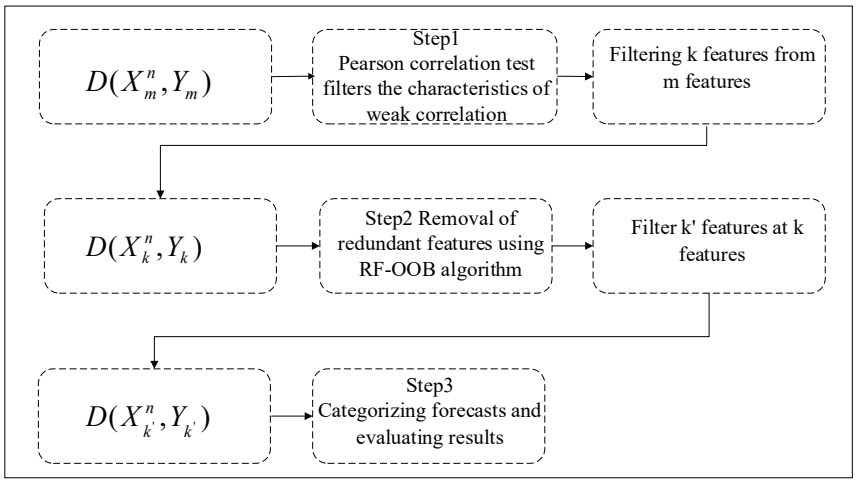

Figure 2. Flowchart of hybrid feature selection.

Using Equation (5), the Pearson correlation coefficients between each feature and the target variables (benign and malignant) are calculated. The feature variables with an absolute value of correlation coefficient greater than or equal to 0.3 are filtered out. Finally, the filtered feature variables are used as candidate features. In order to avoid redundancy among the candidate subsets, it is necessary to carry out the secondary feature screening for RF-OOB on the candidate subsets, and the specific steps are shown in Figure 3: (1) firstly, RF-OOB is applied to calculate the feature importance of each feature, and the features are ranked according to feature importance. The subset of features with the highest feature importance is used as the initial feature combination, and RF-OOB is applied to calculate the model score. (2) Add the subset of features with the second highest feature importance as a new feature combination to be input into the RF-OOB algorithm and calculate the new model score. (3) Add a subset of features one by one according to their importance as a new combination of features and compute a new model classification score. This is iterated until all the feature subsets are traversed. Finally, the feature combination with the highest model classification score is selected as the optimal feature.

5.2. Logistic Regression Optimization Model Based on Batch Gradient Descent (BGD-LR)

The smaller loss function represents the better prediction effect of the model. In order to solve the problem of the poor classification effect of a traditional logistic regression model on the WDBC dataset, this paper uses the BGD algorithm to optimize the LR model so that the loss function reaches the minimum value. The specific steps of Algorithm 1 are as follows:

Algorithm 1 BGD-LR algorithm

Input : a dataset $D\left(X_{k'}^n, Y_{k'}\right)$ that has been filtered with a mixture of features, initialize the θ
Output: prediction results
1. Take the partial derivative of the loss function $J(\theta)$ and compute the gradient using the full training set of samples.
2. Update the model parameters θ according to Equation (10).
3. Repeat steps 2 through 3 for multiple iterations until the specified number of iterations is reached and return θ.
4. Calculate the predicted classification results: calculate the predicted values according to the updated θ and Formula (8) in step 2, and output the classification results.

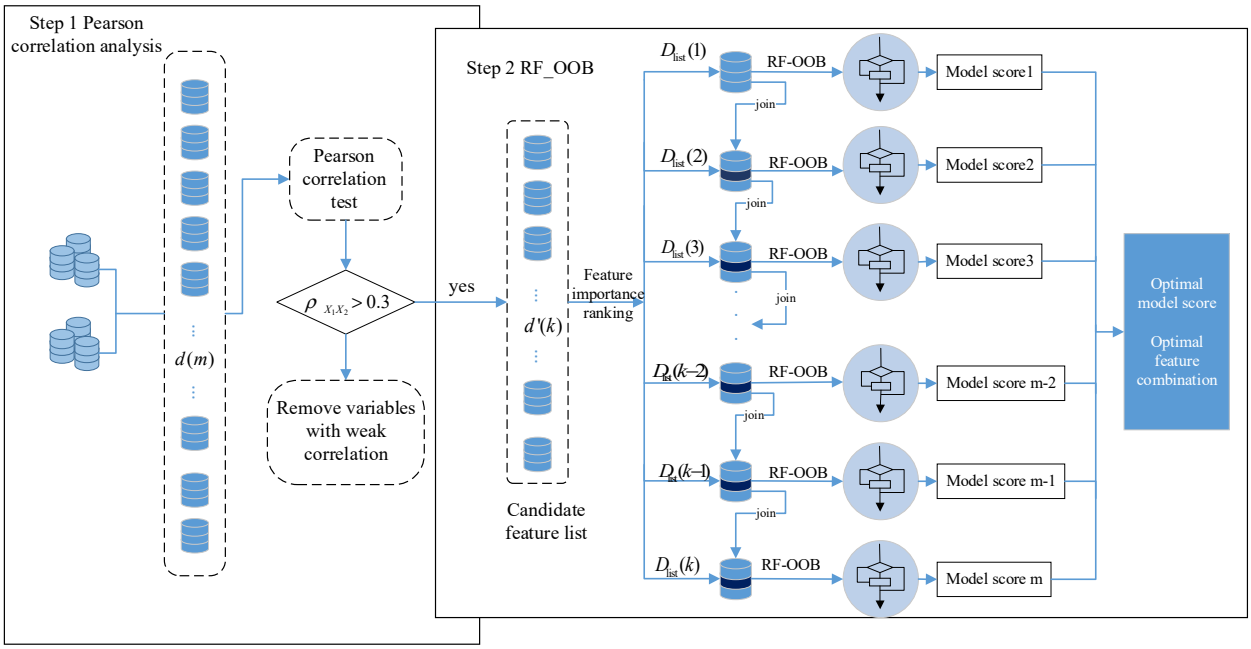

Figure 3. Specific steps of hybrid feature selection: In the first step, the Pearson correlation coefficient method is used. In the second step, the RF-OOB algorithm is used.

5.3. Logistic Regression Optimization Model for Batch Gradient Descent with Differential Privacy (BDP-LR)

In order to solve the problem that traditional LR cannot protect data privacy, this paper uses the BGD algorithm to optimize the loss function of the LR model. At the same time, Gaussian noise is added to each layer of gradient descent, which enables the model to have accurate classification performance while protecting data privacy.

Adding Gaussian noise to the BGD-LR is the core idea of the BDP-LR algorithm. Since the loss function of the LR model is Lipschitz continuous and bounded [49], it means that the global sensitivities of these gradient functions are all bounded.

$$if \left\| x^{(i)} \right\|_2 \leq b \; then \left\| \nabla \ell \left(\theta; x^{(i)}, y^{(i)} \right) \right\|_2 \leq b \tag{11}$$

Thus, for a BGD-LR model, if it can be guaranteed that this gradient is bounded, it can be straightforward to increase the noise by giving the sensitivity as b, with an upper bound on the sensitivity of L_2 obtained by the cropping technique b. The sensitivity formula for the cropping gradient in this paper is:

$$\left\| L_{2\_clip} \left(\nabla \ell \left(\theta; x^{(i)}, y^{(i)} \right), b \right) - L_{2\_clip} \left(\nabla \ell \left(\theta; x^{(i)}, y^{(i)} \right), 0 \right) \right\|_2 \tag{12}$$

The BGD model after adding noise [50] is

$$\theta_{n+1} = \theta_n - \nabla \ell \left(\theta; x^{(i)}, y^{(i)} \right) + \mathrm{N} \left(\sigma^2 \right) \tag{13}$$

The specific steps of Algorithm 2 are as follows:

Algorithm 2 BDP-LR algorithm

Input : a dataset $D\left(X_{k'}^{n}, Y_{k'}\right)$ that has been filtered with a mixture of features, initialize the θ , the privacy budget ε
Output: prediction results
1. Take the partial derivative of the loss function $J(\theta)$ and compute the gradient using the full training set of samples.
2. Add Gaussian noise to the gradient descent algorithm for a single layer: choose a suitable privacy budget ε. According to Equations (11) and (12), the L_2 sensitivity upper bound b obtained by the cropping technique is utilized to increase the Gaussian noise. The θ after adding noise is obtained according to Equation (13).
3. Add Gaussian noise to the BGD algorithm:
a. Based on step 1, add noise to k gradients, and sum the noise gradient values.
b. Calculate the Gaussian noise count value for the number of training samples (sensitivity of 1).
c. Divide the value of the noise gradient in (a) by the value of the Gaussian noise figure in (b).
4. Calculate the predicted classification results: calculate the predicted values according to the updated θ and Formula (8) in step 2, and output the classification results.

5.4. Privacy Analysis of the BDP-LR Algorithm

A and A' are two neighboring datasets, $F(A)$ and $F(A')$ denote the set of all outputs of the neighboring datasets after the BDP-LR algorithm, respectively. S denotes all outputs of the algorithm, then:

$$\begin{aligned}\frac{\Pr[F(A)=S]}{\Pr[F(A')=S]} &= \frac{\Pr\left[f(A)+N\left(\sigma^2\right)=S\right]}{\Pr[f(A\prime)+N(\sigma^2)=S]} = \frac{\Pr\left[N\left(\sigma^2\right)=S-f(A)\right]}{\Pr[N(\sigma^2)=S-f(A')]} \\ &= \frac{\exp(-\frac{[S-f(A)]^2}{2\sigma^2})}{\exp(-\frac{[S-f(A')]^2}{2\sigma^2})} = \frac{\exp(-\frac{[S-f(A)]^2}{2\sigma^2})}{\exp(-\frac{[S-f(A')]^2}{2\sigma^2})} \\ &= \exp\left(\frac{[S-f(A')]^2}{2\sigma^2}-\frac{[S-f(A)]^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{1}{2\sigma^2}\left\{[S-f(A)+\Delta f]^2-[S-f(A)]^2\right\}\right) \\ &= \exp\left(\frac{1}{2\sigma^2}\left[\left(s+\Delta f\right)^2-s^2\right]\right)\end{aligned}$$

Due to the constant positivity of the probability values,

$$\begin{aligned}\left|\ln e^{\frac{1}{2\sigma^2}[(s+\Delta f)^2-s^2]}\right| &= \left|\frac{1}{2\sigma^2}\left[\left(s+\Delta f\right)^2-s^2\right]\right| \\ &= \left|\frac{1}{2\sigma^2}\left(2s\Delta f+\Delta^2 f\right)\right|\end{aligned}$$

When $\sigma^2 \geq c\Delta^2 f/\varepsilon^2$, $c^2 \geq 2\ln(1.25/\delta)$, it can be proven that $\frac{1}{2\sigma^2}\left(2s\Delta f+\Delta^2 f\right) \leq \varepsilon$, then $\frac{\Pr[F(A)=S]}{\Pr[F(A')=S]} \leq e^{\varepsilon}$.

In the batch gradient algorithm, a Gaussian noise mechanism with a privacy budget of ε is executed on each disjoint gradient according to the parallelism and combinatoriality [51] of differential privacy, so that the algorithm satisfies (ε,δ)-differential privacy in each round of iterations.

6. Experimental Results and Analysis

In order to find the best combination of hyperparameters for each model to optimize the breast cancer classification results, in this paper, the grid search method and 5-fold cross-validation method are used to determine the optimal hyperparameters of the models. Then, this paper describes six sets of experiments: the first set of experiments is to standardize

the data by the Z-score standardization; the second set of experiments is to perform hybrid feature selection on the data; the third set of experiments is to verify the effectiveness of the BGD algorithm for LR model optimization; the fourth set of experiments is to verify the effect of hybrid feature selection on the model performance; the fifth set of experiments is to compare the experimental results of the BGD-LR model with the experimental results of other papers without considering privacy protection; and the sixth set of experiments compares the performance of the BDP-LR model with other differential privacy-based machine learning models while considering privacy preservation. Finally, the results are analyzed.

6.1. Experimental Environment and Model Hyperparameters

The operating system used for the experiments is Windows 11, the environment is Python 3.9.7, the processor is Intel(R) Core(TM) m3-6Y30 CPU @ 0.90GHz1.51 GHz, and the RAM is 4.00 GB. The experiments are conducted using the WDBC dataset, and the data are divided into a test set and training set according to a ratio of 8:2. Both the grid search method and the cross-validation method are used to improve the accuracy of the model. The grid search method is a parameter-tuning method to find the best combination of hyperparameters by trying all possible combinations of hyperparameters to improve the accuracy of the model. Also, to avoid model overfitting, the cross-validation method is used to assess the generalization ability of the model. This impact of the differences between the training set and the test set is reduced. The commonly used cross-validation methods are K-fold cross-validation and leave-one-out cross-validation. In this paper, we use 5-fold cross-validation to divide the dataset into five copies. Each time, four copies are used as the training set and the remaining one as the validation set. This is repeated five times, and the average value is taken as the final result. The specific steps are as follows:

1. Firstly, select a set of parameter value ranges for each hyperparameter.
2. Then, evaluate the performance of the adjusted model by the cross-validation method.
3. Finally, select the parameter with the best performance as the best combination.

The optimal parameter combinations of the model determined according to the above method are shown in Table 3.

Table 3. Optimal hyperparameters of the model.

Model	Hyperparameter	Meaning	Value
RF	n_estimators	Number of weak classifiers	200
	oob_score	Whether to use out-of-bag samples	TRUE
LR	penalty	Penalty term	L2
	solver	Optimization algorithm	liblinear
	C	The inverse of the regularized intensity	1
GDP-EBM	min_samples_leaf	Minimum number of samples at leaf nodes	2
	learning_rate	Learning rate	0.03
DP-NB	var_smoothing	Smoothing parameter	0.000000001
DP-DT	max_depth	Maximum number of layers generated	9
DP-RF	max_depth	Maximum number of layers generated	10
	n_estimators	Number of weak classifiers	100

6.2. Experimental Design

In this paper, six groups of experiments are designed, from which the average of the results of one hundred experiments is taken as the final results for the model with added differential privacy due to the randomness of the added noise, and the six groups of experiments are as follows:

1. In order to reduce the influence of data magnitude on the model impact, the data are subjected to Z-score standardization in this paper.

2. In order to eliminate weakly correlated variables and redundant features from breast cancer data, this paper includes hybrid feature selection on the data.
3. In order to test the optimization effect of the BGD model on the LR algorithm, the loss function graph of the BGD-LR is analyzed in this paper.
4. In order to testify to the impact of hybrid feature selection algorithms on model performance, we set up a control group experiment and analyze the results from the four main evaluation indicators: accuracy, precision, recall, and F1-score.
5. The experimental results of this paper are compared with those of other papers. The breast cancer classification model proposed in this paper is compared with existing research results without considering privacy protection.
6. The prediction results of the BDP-LR model in this paper are compared with other machine learning models based on differential privacy when privacy protection is considered.

6.3. Analysis of Experimental Results

6.3.1. Results of Data Standardization

In order to reduce the influence of data magnitude on the model impact, Z-score standardization is applied to the WDBC dataset, and some of the results are shown in Table 4.

Table 4. Results of data standardization.

Compactness_Mean	Concavity_Mean	Concave Points_Mean	Radius_se	Perimeter_se
3.283515	2.652874	2.532475	2.489734	2.833031
−0.48707	−0.02385	0.548144	0.499255	0.263327
1.052926	1.363478	2.037231	1.228676	0.850928
3.402909	1.915897	1.451707	0.326373	0.286593
0.53934	1.371011	1.428493	1.270543	1.273189
1.244335	0.866302	0.824656	−0.25507	−0.3213
0.088295	0.300072	0.646935	0.149883	0.15541

6.3.2. Results of Hybrid Feature Selection

First, the Pearson correlation coefficients between each feature and the target variables are calculated, and values with absolute values of the correlation coefficients greater than or equal to 0.3 are filtered out. Then, the feature importance of the candidate subset is computed using the out-of-bag estimation random forest algorithm. In the meantime, the individual features are ranked according to their importance from highest to lowest, and the final results are shown in Table 5, where $\rho_{X_1X_2}$ represents the Pearson correlation coefficient and number is the sorted sequence number.

According to Table 5, there are 23 candidate features with Pearson correlation coefficient greater than 0.3. The 23 features are screened using the iterative RF-OOB algorithm, and the results of each feature combination are shown in Table 6. The model classification score of the optimal feature combination is 0.96837, and there are 17 features in the feature combination. Therefore, these 17 features are used as the final input values of the breast cancer classification model.

Table 5. The features are ranked in order of feature importance.

Number	Feature	$\rho_{X_1X_2}$	Feature_Importance
0	radius_worst	0.77645	0.15864
1	perimeter_worst	0.78291	0.15337
2	concave_points_worst	0.79357	0.11731
3	area_worst	0.73383	0.11674
4	concave points_mean	0.77661	0.06276
5	area_mean	0.70898	0.05056
6	perimeter_mean	0.74264	0.04596
7	concavity_mean	0.69636	0.04008
8	area_se	0.54824	0.04007
9	radius_mean	0.73003	0.03839
10	concavity_worst	0.65961	0.03644
11	smoothness_worst	0.42147	0.01698
12	texture_worst	0.45690	0.01536
13	texture_mean	0.41519	0.01475
14	perimeter_se	0.55614	0.01408
15	radius_se	0.56713	0.01369
16	symmetry_worst	0.41629	0.01368
17	compactness_worst	0.59100	0.01291
18	compactness_mean	0.59653	0.00994
19	concave points_se	0.40804	0.00884
20	fractal_dimension_worst	0.32387	0.00851
21	smoothness_mean	0.35856	0.00646
22	symmetry_mean	0.33050	0.00447

Table 6. The oob-score for different combinations of features.

Feature Combination	oob-Score
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17, 14, 1, 10, 9, 21	0.96837
13, 15, 20, 16, 7	0.96662
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19	0.96662
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17, 14, 1, 10, 9, 21, 18, 5, 12, 22, 4, 8	0.96639
13, 15	0.96488
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17, 14, 1, 10, 9	0.96488
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17, 14, 1, 10, 9, 21, 18, 5, 12, 22	0.96488
13	0.96487
13, 15, 20, 16	0.96487
13, 15, 20, 16, 7, 3, 2, 6	0.96487
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17, 14, 1, 10, 9, 21, 18, 5, 12, 22, 4	0.96487
13, 15, 20	0.96313
13, 15, 20, 16, 7, 3, 2, 6, 11	0.96310
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17, 14, 1, 10, 9, 21, 18	0.96310
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17	0.96136
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17, 14, 1	0.96136
13, 15, 20, 16, 7, 3	0.95960
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17, 14	0.95960
13, 15, 20, 16, 7, 3, 2, 6, 11, 0	0.95959
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17, 14, 1, 10	0.95959
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17, 14, 1, 10, 9, 21, 18, 5, 12	0.95959
13, 15, 20, 16, 7, 3, 2, 6, 11, 0, 19, 17, 14, 1, 10, 9, 21, 18, 5	0.95785
13, 15, 20, 16, 7, 3, 2	0.95609

6.3.3. Loss Function of the BGD-LR Model

The loss function for training the LR model using the BGD algorithm is shown in Figure 4.

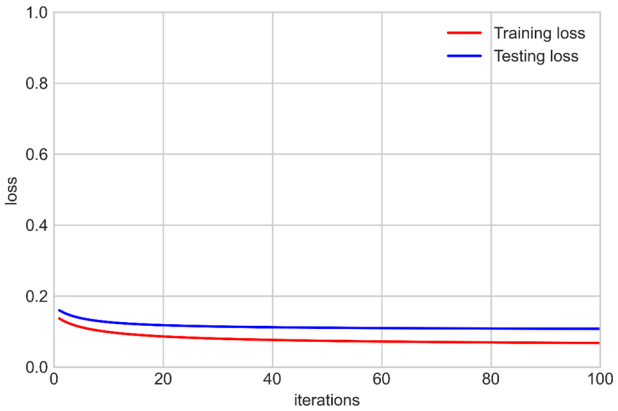

Figure 4. Loss function in the BGD-LR model.

From Figure 4, it can be seen that the loss function values of the training and test sets decrease with each iteration of gradient descent, and with the increase in the number of iterations, the loss function values gradually converge to reach their respective minimum values. Since the loss function values of the train and test sets are low and the difference is between 0.023 and 0.0402, the difference is small. Therefore, from the point of view of the loss function, it can be seen that the difference between the prediction results of the BGD-LR model and the real labels is relatively minor, in other words, the fitting effect is great.

6.3.4. Impact of Hybrid Feature Selection Algorithms on Model Performance

In order to verify the effectiveness of hybrid feature selection, we set up a control group and an experimental group for comparison. The data in the control group are only screened by the Pearson correlation coefficient in feature selection, followed by prediction. The data in the experimental group are screened by the hybrid feature selection method proposed in this paper, and then the prediction of breast cancer is carried out. The comparison results are shown in Table 7.

Table 7. Comparison of results before and after hybrid feature selection.

	Model	Privacy	Accuracy	Precision	Recall	F1-Score
Control group	BGD-LR	0	0.9649	1.0000	0.9545	0.9767
	BDP-LR	0.2	0.8936	0.9840	0.8765	0.9256
		0.4	0.9249	0.9903	0.9117	0.9490
		0.6	0.9454	0.9928	0.9361	0.9634
		0.8	0.9563	0.9949	0.9483	0.9709
		1	0.9566	0.9954	0.9482	0.9711
Experimental group	BGD-LR	**0**	**0.9912**	**1.0000**	**0.9886**	**0.9943**
	BDP-LR	**0.2**	**0.9170**	**0.9849**	**0.9065**	**0.9431**
		0.4	**0.9561**	**0.9933**	**0.9495**	**0.9706**
		0.6	**0.9629**	**0.9959**	**0.9559**	**0.9753**
		0.8	**0.9721**	**0.9975**	**0.9664**	**0.9816**
		1	**0.9777**	**0.9981**	**0.9731**	**0.9853**

According to Table 7 and Figure 5, it can be seen that, compared with the control group, the accuracy, recall, and Fl-score of the experimental group of the BGD-LR model improved by 2.63%, 3.41%, and 1.76%, respectively. For the BDP-LR model, when the privacy budget ε added in each iteration is 0.2, the four main evaluation indicators of the

model increase by 2.34%, 0.09%, 3.00%, and 1.75%, respectively. When the privacy budget ε is 0.4, the four main evaluation indicators of the model improve by 3.12%, 0.3%, 3.78%, and 2.16%, respectively. When the privacy budget ε is 0.6, the indicators of the model increase by 1.75%, 0.31%, 1.98%, and 1.19%, respectively. When the privacy budget ε is 0.8, the four main evaluation indicators improve by 1.58%, 0.26%, 1.81%, and 1.07%, respectively. And when the privacy budget ε is 1, the four main evaluation indicators increase by 2.11%, 0.27%, 2.49%, and 1.42%, respectively. Obviously, after the breast cancer data are processed by hybrid feature selection, the model classification results of the experimental group are all better than those of the control group, therefore, hybrid feature selection can effectively improve the classification performance of the model.

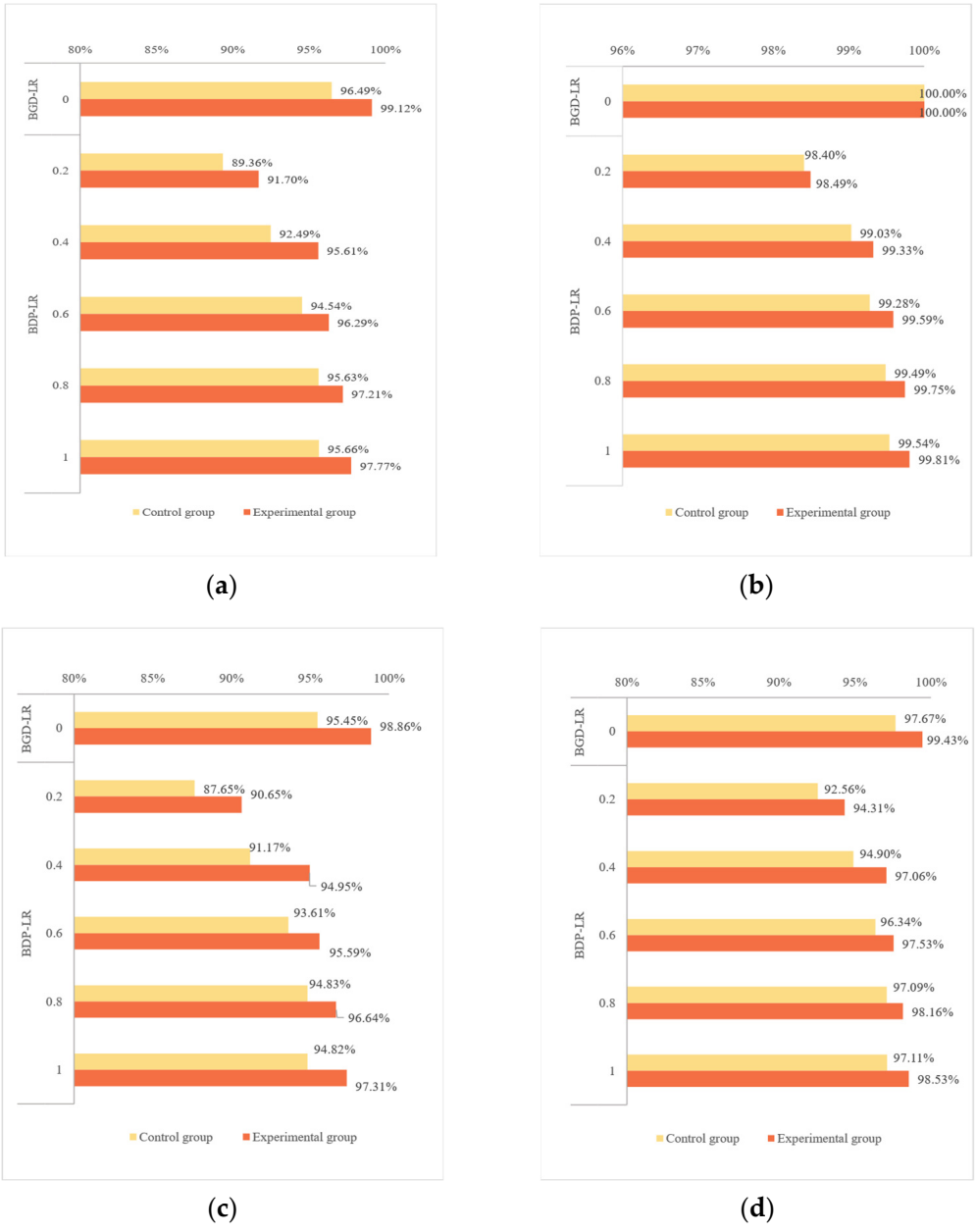

Figure 5. Comparison of experimental results for control and experimental groups: (**a**) accuracy; (**b**) precision; (**c**) recall; (**d**) *F*1-score.

After comparative analysis, it can be seen that the optimal accuracy, precision, recall, and *F*1-score of the BGD-LR model are 0.9912, 1, 0.9886, and 0.9943, respectively. When the added privacy budget ε is 1, the BDP-LR model has the best combined classification

results, and its accuracy, precision, recall, and Fl-score are 0.9777, 0.9981, 0.9731, and 0.9853, respectively.

6.3.5. Comparative Analysis with Previous Studies

In order to further verify the effectiveness of the breast cancer classification model developed in this paper, the classification results of this paper are compared with those of other studies. Firstly, the breast cancer classification model proposed in this paper is compared with the results of existing studies without considering privacy protection, and the results are shown in Table 8.

Table 8. Comparison of the results of the BGD-LR model with other breast cancer classification models.

Literature	Method of Feature Selection	Method of Classification	Time	Accuracy
[26]	ABC	XGBoost	2019	0.928
[21]	GA	SVM	2020	0.988
[20]	GeFeS	KNN	2020	0.985
[18]	χ^2 test + (ET) + (RFE) + RF	ET	2020	0.952
[19]	WCHI2	KNN	2020	0.986
[24]	ALO	BPNN	2020	0.9842
[28]	GWO	KNN	2020	0.948
[23]	BBA	OGCNN	2020	0.935
[22]	Krill herd (KH) + SVM	BPNN	2021	0.978
[27]	Forward selection	LR	2022	0.982
[30]	ESO	RF	2023	**0.9896**
[12]	-	SV-naïve Bayes-3-MetaClassifiers	2020	0.981
[13]	-	IRFRE	2020	0.951
[11]	-	(SVM + LR + NB + DT) + ANN	2022	0.9883
[9]	-	LR	2020	0.981
[8]	-	RF	2022	0.9624
[10]	-	EL	2022	0.9814
this paper	Pearson + RF-OOB	BGD + LR		**0.9912**

GA = genetic algorithm, GeFeS = generalized wrapper-based feature selection, ET = extra tree, RFE = recursive feature elimination, BPNN = back propagation neural network, IRFRE = improved random forest-based rule extraction, BBA = binary bat algorithm, OGCNN = one-pass generalized classifier neural network, EL = ensemble learning technique, WCHI2 = with chi-square feature selection technique, ALO = ant lion optimization algorithm.

As can be seen from Table 8, through comparative analysis, the prediction method for breast cancer classification proposed in this paper outperforms previous research results with an accuracy of 0.9912. Therefore, the hybrid feature selection method and the BGD-LR model used in this paper provide the best classification results.

6.3.6. Comparative Analysis of BDP-LR Model Results with Other Models

The prediction effect of the BDP-LR model in this paper is compared with the DP-NB [35], DP-RF [36], DP-DT [36], and GDP-EBM [37] models under the consideration of privacy preservation. The variation of the four main evaluation indicators with ε for each model is shown in Figure 6.

The results show that when increasing the value of privacy budget ε from 0.001 to 2 with WDBC data, the four main evaluation indicators of each model gradually increase and fluctuate up and down a certain value range with the increase in privacy budget ε. According to Formula (2), the smaller ε is, the better the privacy protection effect is. So, the critical value of ε needs to be selected to provide balance between the model's classification performance and privacy protection effect. According to the trend of the four main evaluation indicators in Figure 6, 0.8 is chosen as the privacy budget value of the BDP-LR model when the model has better classification performance and a stronger privacy protection effect.

At a privacy budget ε of 0.8, the experimental results of the BDP-LR model are compared with other machine learning models based on differential privacy. The performance of each model is evaluated using the ROC curve with AUC, as shown in Figure 7. In addition, the average values of the four main evaluation indicators obtained by running 100 experiments are shown in Table 9.

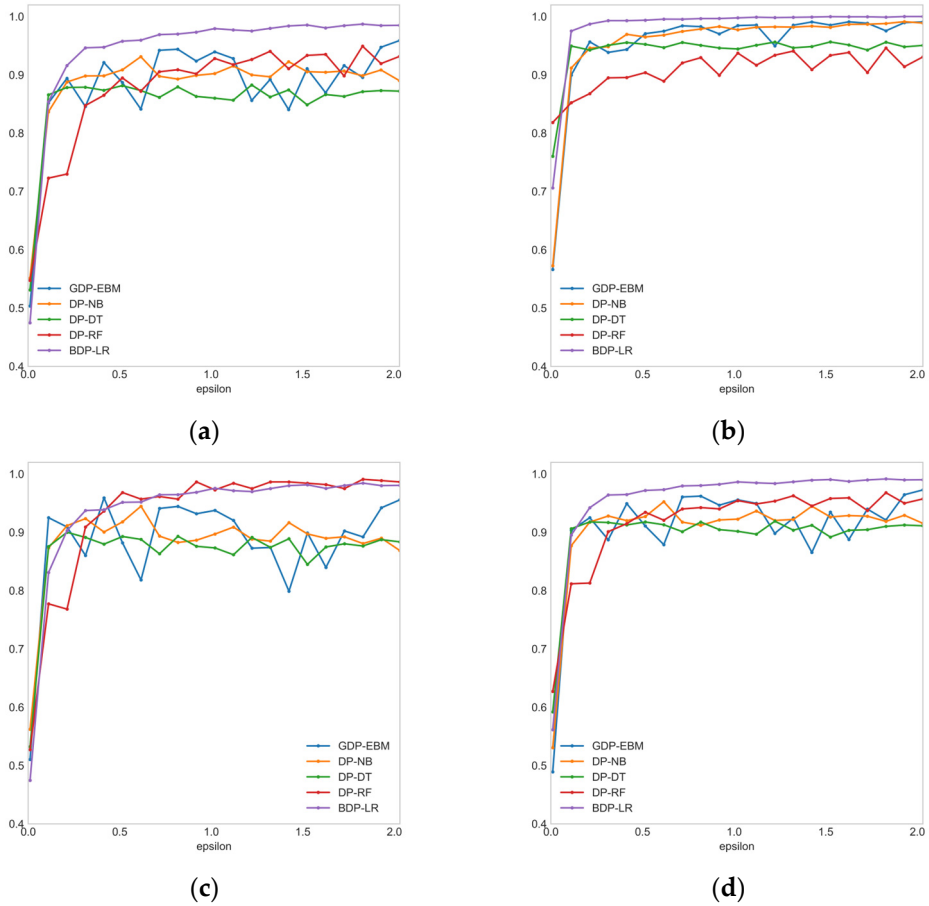

Figure 6. Comparison of the results of the BDP-LR model with other machine learning models based on differential privacy: (**a**) accuracy; (**b**) precision; (**c**) recall; (**d**) F1-score.

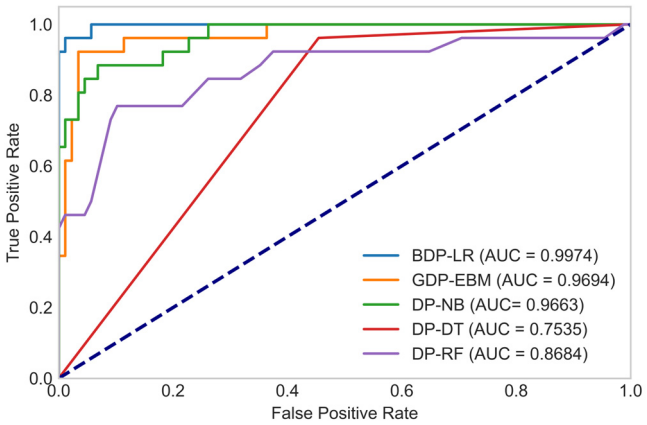

Figure 7. ROC curve for each model when ε is 0.8: the closer the AUC value is to 1, the higher the prediction accuracy is.

Table 9. Classification effect of BDP-LR model compared with other models when ε is 0.8.

Evaluation Indicators	BDP-LR	GDP-EBM	DP-NB	DP-RF	DP-DT
Accuracy	0.9721	0.9439	0.8927	0.9070	0.8793
Precision	0.9975	0.9826	0.9786	0.9276	0.9506
Recall	0.9664	0.9443	0.8825	0.9545	0.8931
*F*1-score	0.9816	0.9620	0.9119	0.9402	0.9175

As shown in Figure 7, the AUC value of the BDP-LR model is 0.9974. The AUC values for the GDP-EBM and DP-NB models are 0.9694 and 0.9663, respectively. The AUC value of the DP-DT model is 0.7535, and the AUC value of the DP-RF model is 0.8684. The higher the prediction accuracy, the closer the AUC value is to 1. Therefore, based on the ROC curves, it can be seen that the BDP-LR model has the highest classification accuracy, followed by the GDP-EBM and DP-NB models.

The experimental results show that the four main evaluation indicators of the BDP-LR model are better than those of the other models when the privacy budget is 0.8, and the four main evaluation indicators are 0.9721, 0.9975, 0.9664, and 0.9816, respectively. Therefore, the logistic regression optimization model based on hybrid feature selection and differential privacy proposed in this paper not only provides high privacy to protect the patients' privacy but also provides superior classification results.

7. Conclusions

Early diagnosis of breast cancer is significant. Applying machine learning to the prediction of breast cancer cells can assist doctors in reducing the rate of leakage and misdiagnosis. However, at this stage, there are still problems of low correct prediction rate and patient privacy leakage. In order to improve the correct rate of breast cancer diagnosis, this paper proposes a breast cancer classification method with higher classification performance, which firstly combines the Pearson correlation test and the RF-OOB algorithm to construct a new hybrid feature selection strategy and secondly optimizes the LR model by using the BGD algorithm. In order to make the model have the effect of protecting patients' privacy, Gaussian noise is added to the BGD algorithm to build the BDP-LR model. In the paper, the accuracy, precision, recall, and F1-score are selected as the four main evaluation indicators of the models. The hyperparameters of each model are determined using the grid search method and the cross-validation method. Experiments on the WDBC dataset show that the hybrid feature selection method proposed in this paper can improve the prediction performance of each model. Comparative analysis shows that the BGD-LR and BDP-LR models constructed in this paper are better. However, the hybrid feature selection method used in this paper has a long computation time, and this paper is limited to combining differential privacy techniques with machine learning models. In the future, further research will be carried out on local differential privacy techniques, deep learning, and so on. At the same time, these studies will be applied to the classification and prediction of breast cancer, contributing to the early diagnosis of breast cancer and the protection of patients' privacy.

Author Contributions: Conceptualization, H.C. and N.W.; methodology, H.C. and N.W.; software, N.W.; validation, N.W. and Y.Z.; formal analysis, N.W.; investigation, N.W.; resources, H.C.; data curation, N.W.; writing—original draft preparation, N.W.; writing—review and editing, H.C., Y.Z., K.M., M.T. and G.C.; supervision, H.C. and G.C.; project administration, H.C.; funding acquisition, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China, grant number 61502156, in part by the teaching and research project of Hubei Provincial Department of Education, grant number 282, and in part by the doctoral startup fund of Hubei University of Technology, grant number BSQD13051.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the fndings of this study are openly available in UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, Y.; Yan, Q.; Fan, C.; Mo, Y.; Wang, Y.; Li, X.; Liao, Q.; Guo, C.; Li, G.; Zeng, Z.; et al. Overview and countermeasures of cancer burden in China. *Sci. China Life Sci.* **2023**, *66*, 1–12. [CrossRef]
2. Jakkaladiki, S.P.; Maly, F. An efficient transfer learning based cross model classification (TLBCM) technique for the prediction of breast cancer. *PeerJ Comput. Sci.* **2023**, *9*, e1281. [CrossRef] [PubMed]
3. Chen, H.; Wang, N.; Du, X.; Mei, K.; Zhou, Y.; Cai, G. Classification Prediction of Breast Cancer Based on Machine Learning. *Comput. Intell. Neurosci.* **2023**, *2023*, 6530719. [CrossRef] [PubMed]
4. Xiao, X. A Study of the Correlation between the Pathologic, Ultrasound, and MRI Manifestations of Breast Cancer and Localized Intravascular Cancerous Emboli. Master's Thesis, University of South China, Hengyang, China, 2021.
5. Qin, J.; Wang, T.Y.; Willmann, J.K. Sonoporation: Applications for Cancer Therapy. *Adv. Exp. Med. Biol.* **2016**, *880*, 263–291.
6. Alromema, N.; Syed, A.H.; Khan, T. A Hybrid Machine Learning Approach to Screen Optimal Predictors for the Classification of Primary Breast Tumors from Gene Expression Microarray Data. *Diagnostics* **2023**, *13*, 708. [CrossRef] [PubMed]
7. Amorim, J.P.; Abreu, P.H.; Fernández, A.; Reyes, M.; Santos, J.; Abreu, M.H. Interpreting Deep Machine Learning Models: An Easy Guide for Oncologists. *Rev. Biomed. Eng.* **2023**, *16*, 192–207. [CrossRef]
8. Arpit, B.; Harshit, B.; Aditi, S.; Ziya, U.; Maneesha, S.; Wubshet, I. Tree-Based and Machine Learning Algorithm Analysis for Breast Cancer Classification. *Comput. Intell. Neurosci.* **2022**, *2022*, 6715406.
9. Ak, M.F. A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. *Healthcare* **2020**, *8*, 111. [CrossRef]
10. Mahesh, T.R.; Vinoth Kumar, V.; Vivek, V.; Karthick Raghunath, K.M.; Sindhu Madhuri, G. Early predictive model for breast cancer classification using blended ensemble learning. *Int. J. Syst. Assur. Eng. Manag.* **2022**. [CrossRef]
11. Naseem, U.; Rashid, J.; Ali, L.; Kim, J.; Haq, Q.E.U.; Awan, M.J.; Imran, M. An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers. *IEEE Access* **2022**, *10*, 78242–78252. [CrossRef]
12. Abdar, M.; Zomorodi-Moghadam, M.; Zhou, X.; Gururajan, R.; Tao, X.; Barua, P.D.; Gururajan, R. A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recognit. Lett.* **2020**, *132*, 123–131. [CrossRef]
13. Wang, S.; Wang, Y.; Wang, D.; Yin, Y.; Wang, Y.; Jin, Y. An improved random forest-based rule extraction method for breast cancer diagnosis. *Appl. Soft Comput.* **2020**, *86*, 105941. [CrossRef]
14. Wang, H.; Zheng, B.; Yoon, S.W.; Ko, H.S. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *Eur. J. Oper. Res.* **2018**, *267*, 687–699. [CrossRef]
15. Zheng, B.; Yoon, S.; Lam, S.S. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Syst. Appl.* **2014**, *41*, 1476–1482. [CrossRef]
16. Kumar, A.; Sushil, R.; Tiwari, A.K. Classification of Breast Cancer using User-Defined Weighted Ensemble Voting Scheme. In Proceedings of the TENCON 2021—2021 IEEE Region 10 Conference (TENCON), Auckland, New Zealand, 7–10 December 2021; pp. 134–139.
17. Jia, X.S.; Sun, X.; Zhang, X. Breast cancer identification using machine learning. *Math. Probl. Eng.* **2022**, *2022*, 8122895. [CrossRef]
18. Chaurasia, V.; Pal, S. Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer. *SN Comput. Sci.* **2020**, *1*, 270. [CrossRef]
19. Zohaib, M.; Akbari, Y.; Shaima, S.; Adnan, K. Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. *J. Chin. Inst. Eng.* **2020**, *43*, 80–92.
20. Sahebi, G.; Movahedi, P.; Ebrahimi, M.; Pahikkala, T.; Plosila, J.; Tenhunen, H. GeFeS: A generalized wrapper feature selection approach for optimizing classification performance. *Comput. Biol. Med.* **2020**, *125*, 103974. [CrossRef]
21. Agustian, F.; Lubis, M.D.I. Particle Swarm Optimization Feature Selection for Breast Cancer Prediction. In Proceedings of the 8th International Conference on Cyber and IT Service Management (CITSM), Pangkal, Indonesia, 23–24 October 2020.
22. Murugesan, S.; Bhuvaneswaran, R.S.; Khanna Nehemiah, H.; Keerthana Sankari, S.; Nancy Jane, Y. Feature Selection and Classification of Clinical Datasets Using Bioinspired Algorithms and Super Learner. *Comput. Math. Methods Med.* **2021**, *2021*, 6662420. [CrossRef]
23. Naik, A.K.; Kuppili, V.; Edla, D.R. Efficient feature selection using one-pass generalized classifier neural network and binary bat algorithm with a novel fitness function. *Soft Comput.* **2020**, *24*, 4575–4587. [CrossRef]
24. Singh, D.; Singh, B.; Kaur, M. Simultaneous feature weighting and parameter determination of Neural Networks using Ant Lion Optimization for the classification of breast cancer. *Biocybern. Biomed. Eng.* **2020**, *40*, 337–351.
25. Zhang, T.; Zhu, T.; Xiong, P.; Huo, H.; Tari, Z.; Zhou, W. Correlated Differential Privacy: Feature Selection in Machine Learning. *IEEE Trans. Ind. Inform.* **2020**, *16*, 2115–2124. [CrossRef]
26. Rao, H.; Shi, X.; Rodrigue, A.K.; Feng, J.; Xia, Y.; Elhoseny, M.; Yuan, X.; Gu, L. Feature selection based on artificial bee colony and gradient boosting decision tree. *Appl. Soft Comput.* **2019**, *74*, 634–642. [CrossRef]
27. Algherairy, A.; Almattar, W.; Bakri, E.; Albelali, S. The Impact of Feature Selection on Different Machine Learning Models for Breast Cancer Classification. In Proceedings of the 7th International Conference on Data Science and Machine Learning Applications (CDMA), Riyadh, Saudi Arabia, 1–3 March 2022.
28. Abdel-Basset, M.; El-Shahat, D.; El-henawy, I.; De Albuquerque, V.H.C.; Mirjalili, S. A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection. *Expert Syst. Appl.* **2020**, *139*, 112824. [CrossRef]
29. Mahesh, T.R.; Vinoth Kumar, V.; Muthukumaran, V.; Shashikala, H.K.; Swapna, B.; Guluwadi, S. Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer. *J. Sens.* **2022**, *2022*, 4649510. [CrossRef]

30. Singh, L.K.; Khanna, M.; Singh, R. Artificial intelligence based medical decision support system for early and accurate breast cancer prediction. *Adv. Eng. Softw.* **2023**, *175*, 103338. [CrossRef]
31. Ji, S.; Du, T.; Li, J.; Shen, C.; Li, B. A Review of Machine Learning Model Security and Privacy Research. *Softw. J.* **2021**, *32*, 41–67.
32. Chen, H.; Zhou, Y.; Mei, K.; Wang, N.; Cai, G. A New Density Peak Clustering Algorithm with Adaptive Clustering Center Based on Differential Privacy. *IEEE Access* **2023**, *11*, 1418–1431. [CrossRef]
33. Zhao, Y.; Yang, M. A Review of Advances in Differential Privacy Research. *Comput. Sci.* **2023**, *50*, 265–276.
34. Dwork, C. Differential privacy. In Proceedings of the 33rd International Colloquium Automata, Languages and Programming, Venice, Italy, 10–14 July 2006; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4052, pp. 1–12.
35. Vaidya, J.; Shafiq, B.; Basu, A.; Hong, Y. Differentially Private Naive Bayes Classification. In Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA, USA, 17–20 November 2013.
36. Fletcher, S.; Islam, M.Z. Differentially private random decision forests using smooth sensitivity. *Expert Syst. Appl.* **2017**, *78*, 16–31. [CrossRef]
37. Nori, H.; Caruana, R.; Bu, Z.; Shen, J.H.; Kulkarni, J. Accuracy, Interpretability, and Differential Privacy via Explainable Boosting. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021.
38. Shen, Q.; Wu, P. Research Progress on Privacy Preserving Technologies in Big Data Computing Environments. *J. Comput.* **2022**, *45*, 669–701.
39. Dwork, C.; Roth, A. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* **2013**, *9*, 211–407. [CrossRef]
40. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In Proceedings of the 3rd Theory of Cryptography Conference (TCC), New York, NY, USA, 4–7 March 2006.
41. Li, Y.; Feng, Y.; Qian, Q. FDPBoost: Federated differential privacy gradient boosting decision trees. *J. Inf. Secur. Appl.* **2023**, *74*, 103468. [CrossRef]
42. Xinzhou, B. Research on Application Technologies of Differential Privacy in Machine Learning. Master's Thesis, University of Science and Technology of China, Hefei, China, 2022.
43. Liu, Y.; Mu, Y.; Chen, K.; Li, Y.; Guo, J. Daily Activity Feature Selection in Smart Homes Based on Pearson Correlation Coefficient. *Neural Process. Lett.* **2020**, *51*, 1771–1787. [CrossRef]
44. Li, Y.; Chen, H.; Li, Q.; Liu, A. Random forest algorithm based on out-of-packet estimation under differential privacy. *J. Harbin Inst. Technol.* **2021**, *53*, 146–154.
45. Sun, Y.; Lin, W. Application of Gradient Descent to Machine Learning. *J. Suzhou Univ. Sci. Technol. Nat. Sci. Ed.* **2018**, *35*, 26–31.
46. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009.
47. Mangasarian, O.L.; William, H.W. *Cancer Diagnosis via Linear Programming*; University of Wisconsin-Madison Department of Computer Sciences: Madison, WI, USA, 1990.
48. Das, M.K.; Chaudhary, A.; Bryan, A.; Wener, M.H.; Fink, S.L.; Morishima, C. Rapid Screening Evaluation of SARS-CoV-2 IgG Assays Using Z-Scores to Standardize Results. *Emerg. Infect. Dis.* **2020**, *26*, 2501–2503. [CrossRef] [PubMed]
49. Du, Q. An Online Logistic Regression Study Based on Differential Privacy. Master's Thesis, Northwest University, Xi'an, China, 2021.
50. Xie, Y.; Li, P.; Wu, C.; Wu, Q. Differential Privacy Stochastic Gradient Descent with Adaptive Privacy Budget Allocation. In Proceedings of the IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 15–16 January 2021.
51. Kairouz, P.; Oh, S.; Viswanath, P. The Composition Theorem for Differential Privacy. *IEEE Trans. Inf. Theory* **2017**, *63*, 4037–4049. [CrossRef]